MULTIMEDIA UNIVERSITY®

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2017/2018

HPB3011 –Bioinformatics Algorithms II

(All sections / Groups)

24 OCTOBER 2017
9:00 - 11:00 AM
(2 hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 3 pages, including this cover page.
2. You are required to attempt all questions. All questions carry equal marks (10).
3. You may use a calculator.
4. Write all your answers in the Answer Booklet provided.

Question 1

a) Paired-end RNA sequencing has become the norm in transcriptomic analysis. State TWO of the advantages of paired-end over single-end sequencing in RNA quantification workflow. [2 marks]

b) Name two types of microarray gene chips and state the main difference. [2 marks]

c) *Gene expression data in RNA-Seq and microarray are not conclusive without proper data preprocessing.*
Based on the statement above, provide two justifications on the need of thorough data preprocessing in RNA-seq and microarray analysis. [2 marks]

d) State any two databases which sequencing data could be obtained. [2 marks]

e) A scientist is performing reference-based transcriptomic analysis on the RNA-Seq generated data from a **bacteria** species. He is now considering to either use Tophat or burrow-wheeler aligner for reads mapping. What would you recommend and why? [2 marks]

Question 2

a) Use K-means clustering to cluster the data into 2 clusters. Solve it manually. Provide justification for terminating the iterations.

	Gene A	Gene B	Gene C	Gene D	Gene E
Sample 1	2.7	5.7	3.4	7.2	5.1
Sample 2	3.5	2.4	3.4	6.8	4.9

[4 marks]

b) Given the following probabilities, use the Bayes approach to find the probability of having diabetes if the test is positive – P(Diabetes |Test+). [2 marks]

	Diabetes	NO Diabetes
Test +	0.8	0.01
Test -	0.2	0.99

overall probability of diabetes = 0.01

c) State and elaborate the three problems of the Hidden Markov Model. [3 marks]

d) How is overfitting represented in the support vector machine? [1 mark]

Continued .....

Question 3

a) Describe the concept of Chou-Fasman and GOR methods in protein structure predictions. Why is GOR method considered more likely to have higher accuracy of structure prediction compared to Chou-Fasman? [3 marks]

b) What approach applied in Profile network from Heidelberg, Rost and Sander (PHD)? Explain the architecture of the PHD method in predicting protein secondary structure. You may also use diagram to illustrate the program. [5 marks]

c) *SCOP2 and CATH are structure classification databases based on different hierarchy.*
Discuss two important roles of protein structure classification in these databases. [2 marks]

Question 4

a) What are the RNA folding problems? [1 marks]

b) In your opinion, comparing the expression of RNAs and structures, which have more significant effect onto the biological function. Justify your answer based on current knowledge in science. [2 marks]

c) What is the "Free Energy Paradigm"? What is the limitation of the current RNA secondary structure prediction methods? What is the strategy toward better RNA structure prediction software? [4 marks]

d) What is the main advantage of Mutual Information and Covariance over Zucker and Nussinov algorithms in RNA structure prediction? [1 marks]

e) Given the matrix below, draw the structure of the RNA. [2 marks]

	C	C	C	U	U	U	U	A	G	G
C	0	0	0	0	0	0	0	1	2	3
C		0	0	0	0	0	0	1	2	3
C			0	0	0	0	0	1	2	2
U				0	0	0	0	1	1	1
U					0	0	0	1	1	1
U						0	0	1	1	1
U							0	1	1	1
A								0	0	0
G									0	0
G										0

END OF PAPER